DIGITALCOMMONS @WAYNESTATE

Journal of Modern Applied Statistical Methods

Volume 6 | Issue 1 Article 22

5-1-2007

A Spline-Based Lack-of-Fit Test for Independent Variable Effect

Chin-Shang Li
St. Jude Children's Research Hospital, chinshang.li@stjude.org

Wanzhu Tu
Indiana University School of Medicine

Journal of Modern Applied Statistical Methods
May, 2007, Vol. 6, No. 1, 239-247

1538 – 9472/07/$95.00

A Spline-Based Lack-Of-Fit Test for Independent Variable Effect in Poisson Regression

Chin-Shang Li
St. Jude Children's Research Hospital

Wanzhu Tu
Indiana University School of Medicine

In regression analysis of count data, independent variables are often modeled by their linear effects under the assumption of log-linearity. In reality, the validity of such an assumption is rarely tested, and its use is at times unjustifiable. A lack-of-fit test is proposed for the adequacy of a postulated functional form of an independent variable within the framework of semiparametric Poisson regression models based on penalized splines. It offers added flexibility in accommodating the potentially non-loglinear effect of the independent variable. A likelihood ratio test is constructed for the adequacy of the postulated parametric form, for example log-linearity, of the independent variable effect. Simulations indicate that the proposed model performs well, and misspecified parametric model has much reduced power. An example is given.

Key words: B-splines, likelihood ratio test, loglinear model, penalized likelihood, Poisson regression model.

Introduction

The Poisson regression model is among the most frequently used statistical tools in event count analysis. It has been successfully used in numerous applications (e.g. McCullagh & Nelder, 1989; Cameron & Trivedi, 1998). In its parametric form, the model is constructed as follows: Let Y_i be the number of occurrences of an event of interest for the ith subject and x_i be an independent variable, $i = 1\ldots,n$. Assuming that Y_i follows a Poisson distribution, i.e., $Y_i \sim \text{Poisson}(\mu(x_i))$ with mean $\mu(x_i)$, write the Poisson density as follows:

$$f(y_i \mid x_i) = \frac{\exp(-\mu(x_i))\mu(x_i)^{y_i}}{y_i!}, \quad (1.1)$$

where $\mu(x_i) = E[Y_i \mid x_i]$ is the mean function, $i = 1\ldots,n$.

Under this formulation, the model depicts the dependency of the event counts on x_i via a logarithmic link function,

$$\log(\mu(x_i)) = h(x_i, \boldsymbol{\beta}), \quad (1.2)$$

where $h(\cdot;\boldsymbol{\beta})$ is a known functional form apart from the parameter vector $\boldsymbol{\beta}$. In other words, the model assumes a log parametric form for the independent variable effect. Although the theory does not restrict $h(\cdot;\boldsymbol{\beta})$ to a linear form, in practice, however, most analysts choose to use $h(x;\boldsymbol{\beta}) = \beta_0 + \beta_1 x$, which is often referred to as the loglinear model. Despite its popularity, the

Chin-Shang Li is an Associate Member in the Department of Biostatistics at St. Jude Children's Research Hospital, Memphis, Tennessee. Email him at chinshang.li@stjude.org. Wanzhu Tu is an Associate Professor in the Division of Biostatistics at Indiana University School of Medicine, and a research scientist at Indiana University Center for Aging Research and Regenstrief Institute, Indianapolis, Indiana. Email him at wtu1@iupui.edu. This research was supported in part by grants R01 HD042404, HL69399 (W. Tu), and CA21765 (C.S. Li) from the National Institutes of Health and by the American Lebanese Syrian Associated Charities (C.S. Li).

validity of this postulated loglinear form is rarely verified, possibly due to the lack of readily accessible testing procedures. In addition, the consequences of a misspecified functional form of the independent variable are not well studied.

An alternative approach is to replace the linear predictor $h(\cdot;\boldsymbol{\beta})$ with a regression spline. This approach is semiparametric in nature because it has not only a parametric component for the data distribution (Poisson in the case of counts), but also a nonparametric component involving the predictor (e.g. Ruppert et al., 2003). Such an approach is known to enhance the modeling flexibility in regression analysis. Regression spline techniques have been used frequently to estimate independent variable effects in generalized linear models (e.g., Eilers & Marx, 1996). But testing procedures based on regression splines have attracted considerably less attention.

The purpose of this article is to construct a general test for the inference concerning the adequacy of a given functional form of the independent variable effect in count data analysis. The proposed test contributes to the existing literature of count data analysis by providing a practical way for the determination of functional forms of independent variables.

Example

Patients with chronic diseases, such as congestive heart failure (CHF), must take medications regularly to prevent disease exacerbation that requires costly health care services such as emergency department (ED) visits and hospitalization admissions. In a study on medication adherence in older adults with CHF, participants were monitored for their medication use during a 1-year study period. Eligible participants were English-speaking, 50 years of age or older, had a diagnosis of CHF, and were currently prescribed for at least one cardiovascular medication, including angiotensin-converting enzyme inhibitors, angiotensin II-receptor antagonists, β-adrenergic receptor antagonists, digoxin, loop and nonloop diuretics, and an aldosterone antagonist. Upon enrollment, participants were provided electronic medication container lids for their cardiovascular medications. The electronic container lids automatically recorded the dates and times of the lid openings (Tu et al., 2005). Assuming that the patient took the prescribed amount of medication each time the lid was opened, patient's medication adherence to the prescribed drug was calculated as the percentage of dose taken during the observation period according to the prescribed regimen.

For example, if 30 openings were recorded during a 1-month period for a b.i.d. (twice a day) drug, the medication adherence was $r = 30/60 = 50\%$, meaning the patient took only half of the medicine that he was supposed to take. Although 100% is the target level for medication adherence, values that are significantly less or more than 100% would represent suboptimal medication-taking behavior on the part of the patient. Therefore, in pharmacy practice, researchers often calculate the patient's deviation in medication consumption from the target level ($|1-r|$) and report an adjusted adherence $x=1-|1-r|$ as a percentage between 0 and 100%. For a more detailed discussion (see Hope et al., (2004).

When a patient was on multiple cardiovascular drugs, his overall adjusted medication adherence was summarized as the average level of the adjusted medication adherence values for all of the study drugs. An important issue of this study is to understand the relationship between adjusted medication adherence and disease exacerbation. Herein, the number of ED visits during the 1-year follow-up period is used as the primary outcome of interest.

For the purpose of illustration, consider a subset of the study data: 93 subjects who belong to the New York Heart Association (NYHA) Class III. The NYHA classification is one of the most commonly used clinical classification systems for patients with heart failure. Typical NYHA Class III patients experience a marked limitation of physical activities, such as walking one to two blocks on the level or climbing more than one flight of stairs under normal conditions. Patients are comfortable at rest, but more than usual physical activity causes fatigue, palpitation, dyspnea, anginal pain, or a combination thereof.

Methodology

Semiparametric Poisson Model

Using the previously introduced notation, we write the response variables as $Y_i \sim \text{Poisson}(\mu(x_i))$ for $i = 1\ldots, n$. To model the effect of the independent variable x_i on the response variable, link the mean function $\mu(x_i)$ to x_i via $g(\mu(x_i)) = h(x_i)$, where $g(\cdot)$ is a monotone differentiable link function, and $h(\cdot)$ is the predictor function. To differentiate the proposed model from the traditional loglinear predictor, write $g(\mu(x_i)) = s(x_i)$, where $s(x_i)$ is a smooth function to be estimated from the data. Under the commonly used log-link function, there is a semiparametric Poisson regression model:

$$g(\mu(x)) = \log(\mu(x)) = s(x). \qquad (3.1)$$

Equivalently, write $\mu(x) = \exp(s(x))$. Because the B-spline basis is numerically more stable for the representation of a spline function than the truncated-power basis is, to approximate the unknown function $s(\cdot)$, parameterize it with the cubic B-splines as

$$s(x) = \sum_{k=1}^{K} \theta_k B_k(x), \qquad (3.2)$$

where B_k are the cubic B-spline basis functions for s; θ_k are spline coefficients; and $K = q + 4$, where q is the number of knots; see de Boor (1978) for details of computation of B-splines of any degree from B-splines of a lower degree. These knots were chosen to be equally spaced with respect to the quantiles of the distinct values of x_i and set $q = \min[(\text{number of distinct values of } x/6), 30]$, where $[a]$ is the greatest integer less than or equal to a; for reference on the selection of q, (see Ruppert, 2002). Let $\mathbf{B}_{x_i} = (B_1(x_i), \ldots, B_K(x_i))^T$ and $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_K)^T$ and expressing s in vector notation as $s(x_i) = \mathbf{B}_{x_i}^T \boldsymbol{\theta}$, rewrite the Poisson density in (1.1) as

$$f(y_i \mid x_i) = \frac{\exp[-\exp(\mathbf{B}_{x_i}^T \boldsymbol{\theta})]\exp(\mathbf{B}_{x_i}^T \boldsymbol{\theta})^{y_i}}{y_i!}, \qquad (3.3)$$

Parameter Estimation

To estimate $\boldsymbol{\theta}$ and prevent overfitting, employ a penalized likelihood approach with a discrete approximation to the integrated squared second derivative of s, $\int s''(x)^2 dx$, which is used as a measure of its roughness. Therefore, the penalized log-likelihood $\ell(\boldsymbol{\theta}) - (1/2)\int s''(x)^2 dx$ is approximated by

$$\ell_{pen}(\boldsymbol{\theta}) = \ell(\boldsymbol{\theta}) - \frac{1}{2}\lambda \boldsymbol{\theta}^T \mathbf{K} \boldsymbol{\theta}, \qquad (3.4)$$

where the log-likelihood $\ell(\boldsymbol{\theta})$ is

$$\ell(\boldsymbol{\theta}) = \sum_{i=1}^{n} \{y_i \mathbf{B}_{x_i}^T \boldsymbol{\theta} - \exp(\mathbf{B}_{x_i}^T \boldsymbol{\theta}) - \log y_i!\}.$$

λ is a smoothing parameter to be chosen. It is used to govern the tradeoff between goodness-of-fit and smoothness; $\boldsymbol{\theta}^T \mathbf{K} \boldsymbol{\theta} = \sum_{k=1}^{K} (\Delta^2 \theta_k)^2$ for $\Delta^2 \theta_k = \theta_k - 2\theta_{k-1} + \theta_{k-2}$ and $\mathbf{K} = \mathbf{D}_2^T \mathbf{D}_2$ for $\mathbf{D}_2$ being the matrix representation of the difference operator Δ^2. For details of similar operation, see Eilers & Marx (1996). To estimate $\boldsymbol{\theta}$, set the first-order partial derivatives of the penalized log-likelihood in (3.4) with respect to $\boldsymbol{\theta}$ equal to $\mathbf{0}$, $q(\boldsymbol{\theta}) = \partial \ell_{pen}(\boldsymbol{\theta}) / \partial \boldsymbol{\theta} = 0$. Then, solve iteratively the following weighted least-squares equations in matrix notation:

$$(\mathbf{BWB} + \lambda \mathbf{K})\boldsymbol{\theta} = \mathbf{B}^T \mathbf{W} \mathbf{y}^*, \qquad (3.5)$$

where

$$\mathbf{B} = (\mathbf{B}_{x_i}^T, \ldots, \mathbf{B}_{x_i}^T)^T,$$

$$\mathbf{W} = \text{diag}(\exp(\mathbf{B}_{\mathrm{x_1}}^{\mathrm{T}}\boldsymbol{\theta}), \ldots, \exp(\mathbf{B}_{\mathrm{x_n}}^{\mathrm{T}}\boldsymbol{\theta})),$$

and

$$\mathbf{y}^* = (y_1^*, \ldots, y_n^*)^T$$

is the working response vector for

$$\mathrm{y_i^*} = \frac{\mathrm{y_i} - \exp(\mathbf{B}_{\mathrm{x_i}}^{\mathrm{T}}\boldsymbol{\theta})}{\exp(\mathbf{B}_{\mathrm{x_i}}^{\mathrm{T}}\boldsymbol{\theta})} + \mathbf{B}_{\mathrm{x_i}}^{\mathrm{T}}\boldsymbol{\theta}.$$

The concept of the effective number of degrees of freedom (Edf) is used to choose the value of the smoothing parameter λ, $\text{Edf} = \text{tr}(\mathbf{S}_\lambda) - 1$, where

$$\mathbf{S}_\lambda = \mathbf{B}(\mathbf{BWB} + \lambda\mathbf{K})^{-1}\mathbf{B}^T\mathbf{W}$$

is called the smoother matrix (see Hastie & Tibshirani, 1990).

The solution to (3.5), denoted by $\hat{\boldsymbol{\theta}}_{pen}$, is called a maximum penalized likelihood estimate of $\boldsymbol{\theta}$. The cubic B-spline fitted mean function is $\hat{\mu}_{\mathrm{B}} = \exp(\hat{s}_B(x)) = \exp(\mathbf{B}_{\mathrm{x}}^{\mathrm{T}}\hat{\boldsymbol{\theta}}_{pen})$.

Inference

Let $\mathbf{Q}(\boldsymbol{\theta}) = \partial^2 \ell_{pen}(\boldsymbol{\theta}) / \partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T$ be the second-order partial derivatives of the penalized log-likelihood in (3.4) with respect to $\boldsymbol{\theta}$. Then, approximate the covariance matrix of $\hat{\boldsymbol{\theta}}_{pen}$ as $\text{cov}(\hat{\boldsymbol{\theta}}_{pen}) \approx \{-\mathbf{Q}(\boldsymbol{\theta})\}^{-1} \text{var}\{\mathbf{q}(\boldsymbol{\theta})\} \{-\mathbf{Q}(\boldsymbol{\theta})\}^{-1}$ and, hence, estimate it as follows:

$$\hat{\text{cov}}(\hat{\boldsymbol{\theta}}_{pen}) \approx \{-\mathbf{Q}(\hat{\boldsymbol{\theta}})\}^{-1}. \qquad (3.6)$$

From (3.6), obtain a confidence interval for $s(x)$ by computing the estimated variance of $\hat{s}_B(x)$ as follows:

$$\hat{\text{var}}(\hat{s}_B(x)) = \mathbf{B}_x^T \hat{\text{cov}}\{\hat{\boldsymbol{\theta}}_{pen}\}\mathbf{B}_x,$$

where

$$\hat{s}_B(x) = \mathbf{B}_x^T\hat{\boldsymbol{\theta}}_{pen}.$$

Accordingly, a confidence interval for $\exp(s(x))$ can be obtained by using the delta method and the estimated variance of $\hat{s}_B(x)$.

The proposed estimation method is used with deviance, or log-likelihood ratio statistic, to assess the adequacy of a postulated parametric form of the model in (1.2), i.e., testing the following parametric null hypothesis:

$$\mathrm{H}_0 : \log(\mu(\cdot)) = h(\cdot;\boldsymbol{\beta}). \qquad (3.7)$$

Then, test for the lack of fit of the postulated parametric model in (3.7) by using the log-likelihood ratio test (LRT) statistic

$$-2\log\frac{\sup\prod_{i=1}^{n}\exp\{-\exp[h(x_i;\boldsymbol{\beta})]\}\exp[y_i h(x_i;\boldsymbol{\beta})]}{\prod_{i=1}^{n}\exp\{-\exp[\mathbf{B}_{x_i}^T\hat{\boldsymbol{\beta}}_{pen}]\}\exp[y_i\mathbf{B}_{x_i}^T\hat{\boldsymbol{\beta}}_{pen}]} \qquad (3.8)$$

and comparing its value to its asymptotic limiting chi-square distribution with *D* degrees of freedom, where *D* = Edf - the number of the postulated parametric model parameters + 1 (see Hastie & Tibshirani, 1990). This is called a smoothing log-likelihood ratio test (SLRT). By comparing the likelihood of the postulated parametric model $h(\cdot;\boldsymbol{\beta})$ with that of a B-spline model, the SLRT allows for inference on the adequacy of the postulated model. For example, the SLRT can be used to assess the appropriateness of the loglinear effect of x by testing

$$\mathrm{H}_0 : \log(\mu(x)) = \beta_0 + \beta_1 x$$

versus

$$\mathrm{H}_a : \log(\mu(x)) = s(x),$$

where $s(x)$ is a general nonlinear smooth function.

Simulation Study

A Monte Carlo simulation study is conducted to assess the performance of the SLRT. The values of the independent variable are equally spaced design points $x_i = (2i-1)/2n$, $i = 1,\ldots,n = 100$; 1000 data sets were generated for each configuration of the experiments. The goal is to test the adequacy of the following loglinear model:

$$\mathrm{H}_0 : \log(\mu(x;\boldsymbol{\beta})) = \beta_0 + \beta_1 x. \quad (4.1)$$

To assess the empirical power of the SLRT, data were generated from the model in (1.1) with the following logarithmic mean functions:

$$\log(\mu(x;\boldsymbol{\beta},\gamma)) = \beta_0 + \beta_1 x + \gamma \sin(3\pi x), \quad (4.2)$$

where

$$\boldsymbol{\beta} = (\beta_0, \beta_1) = (1.159, -0.675),$$

and

$$\gamma = 0, 0.05, 0.1, \ldots, 0.5.$$

Note that $\gamma = 0$ corresponds to the null hypothesis. In this simulation study, the $\boldsymbol{\beta}$ values were chosen to reflect the estimated coefficient values of the example data, and the value of the smoothing parameter λ was chosen by fixing $\mathrm{Edf} = \mathbf{tr}(\mathbf{S}_\lambda) - 1 = 4$ while estimating the cubic B-spline coefficients. The null model is rejected if the observed value of the SLRT statistic in (3.8) with $h(x;\boldsymbol{\beta},) = \beta_0 + \beta_1 x,$ exceeds the 0.95-quantile of the chi-square distribution with D = 3 degrees of freedom. This is called an asymptotic smoothing log-likelihood ratio test (ASLRT).

An obvious variant of the ASLRT is to approximate the 0.95-quantile of the distribution of the SLRT statistic via 200 Monte Carlo simulations for each sample and reject the null model if the observed value of the SLRT statistic exceeds the approximated 0.95-quantile. This test is called the Monte Carlo smoothing log-likelihood ratio test (MSLRT).

For the purpose of comparative evaluation, it is important to establish a benchmark for the power of the inference when the true model is known. To do so, consider the following parametric likelihood ratio test for $\gamma = 0$ under the assumption that the true model is known:

$$-2\log \frac{\sup \prod_{i=1}^{n} \exp\{-\exp[\beta_0 + \beta_1 x_i]\} \exp[y_i(\beta_0 + \beta_1 x_i)]}{\prod_{i=1}^{n} \exp\{-\exp[\beta_0 + \beta_1 x_i + \gamma \sin(3\pi x_i)]\} \exp[y_i(\beta_0 + \beta_1 x_i + \gamma \sin(3\pi x_i)]}. \quad (4.3)$$

$$-2\log \frac{\sup \prod_{i=1}^{n} \exp\{-\exp[\beta_0 + \beta_1 x_i]\} \exp[y_i(\beta_0 + \beta_1 x_i)]}{\prod_{i=1}^{n} \exp\{-\exp[a_0 + a_1 x_i + a_2 x_i^{\,2}]\} \exp[y_i(a_0 + a_1 x_i + a_2 x_i^{\,2})]}. \quad (4.4)$$

Because this test directly compares the likelihood of the postulated model (4.1) with the true parametric model (4.2) which generates the data, the parametric test in (4.3) is referred to as an exactly specified parametric log-likelihood ratio test (ESPLRT). The null model is rejected if the observed value of the LRT statistic in (4.3) was greater than the 0.95-quantile of the chi-square distribution with one degree of freedom. When the null model is appropriate, the exactly specified test ESPLRT in (4.3) follows an asymptotic chi-square distribution with degree of freedom and should have the best power among all competitors. This asymptotically optimal test, however, provides only a benchmark for the power comparison and is of limited practical values, because its use requires specific knowledge of the form of the true model.

In practical data analysis, when there are clear indications that an independent variable effect is not linear, data analysts often attempt to alleviate the lack-of-fit by including a quadratic term. To assess the potential power loss associated with such practice when a wrong model is used, consider a misspecified quadratic parametric model, $\log(\mu(x)) = a_0 + a_1 x + a_2 x^2$, and use it to construct a parametric likelihood ratio test:

Because the misspecified parametric model $\log(\mu(x)) = a_0 + a_1 x + a_2 x^2$ is used to fit data generated under the true model $\log(\mu(x;\boldsymbol{\beta},\gamma)) = \beta_0 + \beta_1 x + \gamma \sin(3\pi x)$, in (4.2), this parametric test is referred to as a misspecified parametric likelihood ratio test (MSPLRT). By comparing the empirical power of the MSPLRT with that of the benchmark test, ESPLRT, and other competitors, it is possible to assess the power loss when a misspecified quadratic model is used in data analysis.

The empirical powers of the four candidate tests, ESPLRT, MSPLRT, ASLRT, and MSLRT, are depicted in Figure 1. While the empirical significance level of the ASLRT tends to be slightly higher than the nominal level 0.05, its power is the closest to that of the exactly specified parametric likelihood ratio test (ESPLRT), the benchmark test, among all three other candidates. A close second is the Monte Carlo version of the smoothing likelihood ratio test, MSLRT. On the other hand, the misspecified parametric likelihood ratio test MSPLRT suffers a severe loss of power, highlighting the consequences of making inference under misspecified parametric regression models.

Example Data Analysis

The practical use of the proposed semiparametric Poisson regression model and lack-of-fit tests is illustrated by examining the effect of the adjusted medication adherence on ED utilization in a study of 93 NYHA Class III patients. Consider the number of all-cause ED visits as the response variable. All-cause ED visits include ED admissions for any reason. The use of all-cause ED visits in this analysis is justified, because acute exacerbation in patients with heart failure does not always occur in the form of CHF, coronary artery disease, or cardiovascular diseases. Sometimes, it results in complications in other organs, which would be recorded as noncardiovascular-related conditions in the medical records. The adjusted medication adherence to all prescribed cardiovascular medications is the independent variable of interest. Therefore, do not restrict the effect of the overall adjusted medication adherence (x) to be loglinear. Instead, use the proposed inference procedure to test for the loglinear effect of the independent variable x on the number of ED visits. That is, test the following hypotheses:

$$H_0 : \log(\mu(x)) = \beta_0 + \beta_1 x,$$

versus

$$H_a : \log(\mu(x)) = s(x).$$

In this example, $\text{Edf} = \mathbf{tr}(\mathbf{S}_\lambda) - 1 = 4$ is fixed to choose the value of λ for the estimation of cubic B-spline coefficients. By using the smoothing log-likelihood ratio test (SLRT) statistic with an asymptotic chi-square distribution with 3 degrees of freedom, it is found that that the independent variable has a significantly non-loglinear effect on the number of ED visits (ASLRT p-value = 0.029).

The distribution of the SLRT statistic is approximated through 200 Monte Carlo simulations. The p-value from the Monte Carlo based test MSLRT is 0.015, which is the

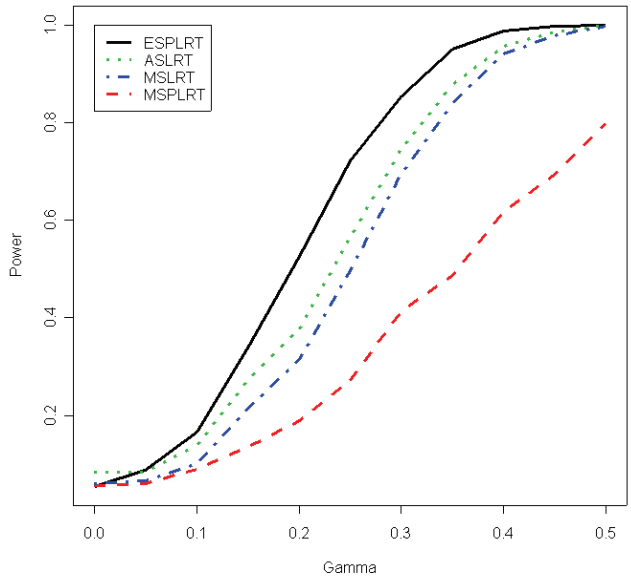

Figure 1: Empirical power curve comparison of the ESPLRT, ASLRT, MSLRT, and MSPLRT for the null model $H_0 : \log(\mu(x;\boldsymbol{\beta})) = \beta_0 + \beta_1 x$ and alternative model $\log(\mu(x;\boldsymbol{\beta},\gamma)) = \beta_0 + \beta_1 x + \gamma \sin(3\pi x)$, where $\boldsymbol{\beta} = (\beta_0, \beta_1)^T = (1.159, -0.675)^T$. Abbreviations: ASLRT, asymptotic smoothing log-likelihood ratio test; ESPLRT, the exactly specified parametric log-likelihood ratio test; MSLRT, Monte Carlo smoothing log-likelihood ratio test; MSPLRT, misspecified parametric log-likelihood ratio test.

proportion of simulated values of the test statistic exceeding the observed value of the SLRT statistic 9.0033. Therefore, the SLRT based on the Monte Carlo simulations again confirmed the lack of log linearity in the medication adherence effect. Figure 2 shows the cubic B-spline fitted mean function $\mu(x)$ and 95% point-wise confidence interval for the mean function and the parametrically fitted mean function $\hat{\mu}(x) = \exp(1.159 - 0.675x)$; Figure 3 further confirms that the cubic B-spline fitted function and 95% point-wise confidence interval for the functional form of the effect of x and the parametrically fitted function $1.159 - 0.675x$.

From the perspective of pharmacotherapy, the lack of log linearity in the effect of medication adherence is perhaps not entirely surprising: underconsumption of cardiovascular drugs often leads to decompensation in patients with CHF, and overdosing can cause dangerous hypotension. Both are likely to result in increased ED use. Because the adherence data are in the adjusted form $(x = 100(1 - |1 - r|)\%)$ where overconsumption of the medication was converted to a percentage less than 100%, a deviation from the target level (100%) could be the results of overconsumption as well as underconsumption. Figures 2 and 3 showed an increase in ED admission when adjusted medication adherence around 0.8, possibly caused by the folding of the raw adherence measure. Therefore, the loglinear relationship forced by the parametric Poisson regression model would not be adequate and the proposed semiparametric model would provide a relief in such a data situation.

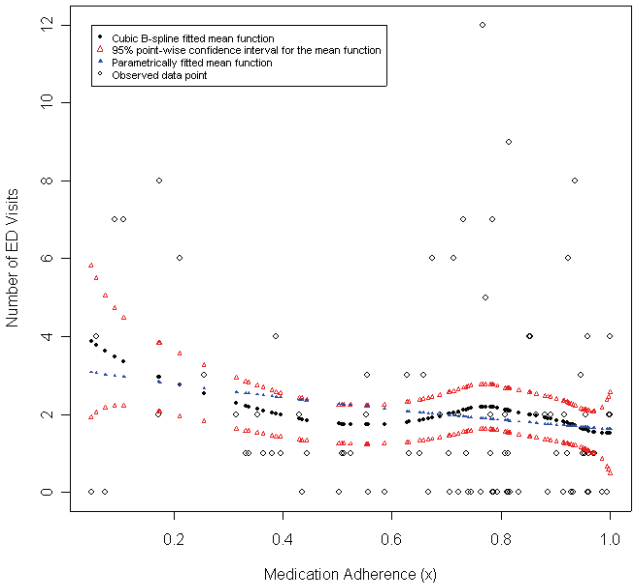

Figure 2: Cubic B-spline fitted mean function $\hat{\mu}_B(x)$ and 95% point-wise confidence interval for the mean function $\mu(x)$ of medication adherence (x) and parametrically fitted mean function $\hat{\mu}(x) = \exp(1.159 - 0.675x)$ for $\mu(x)$.

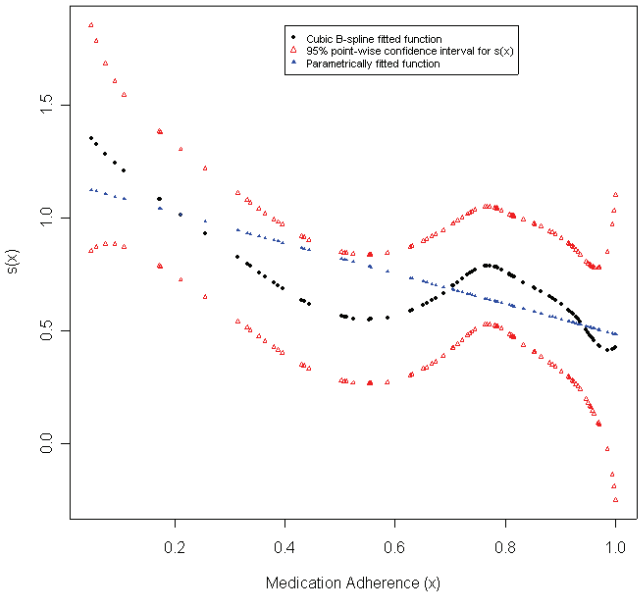

Figure 3: Cubic B-spline fitted function $\hat{s}_B(x)$ and 95% point-wise confidence interval for the functional form $s(x)$ of the effect of medication adherence (x) and parametrically fitted function $1.159 - 0.675x$ for the effect of x.

Conclusion

Cubic B-spline basis functions can be used in Poisson regression analysis in a flexible manner, without imposing any particular functional form on the effects of independent variables. An easily implementable estimation procedure was examined for the coefficients of cubic B-spline basis functions by using a penalized likelihood approach. Fitting B-splines is usually not more difficult than that of a polynomial regression.

Because the selection of the number and locations of the knots is an important issue, it is the topic of much research in nonparametric regression methods (Ruppert, 2002; Lindstrom, 1999). Knots were chosen to be equally spaced with respect to the quantiles of the distinct values of the independent variable and set the number of chosen knots to be $\min\left[\left(\frac{\text{number of distinct values of covariate}}{6}\right),30\right]$ (Ruppert, 2002).

This study has shown that by smoothing the effect of an independent variable, the proposed method allows for a test of the lack of fit of a postulated parametric model by the use of likelihood ratio method. As shown in the simulation study, the proposed test has the ability to detect more general alternative models and is superior to parametric likelihood ratio tests unless the true model is known. This is of great practical importance, because in most real data applications, the true parametric forms of independent variable effects are usually unknown. Therefore, investigators must consider the consequences of statistical inferences under misspecified parametric regression models. Specifically, this simulation study showed that the common practice of adding a quadratic term to the linear predictor could severely undercut the power of inference.

The scope was restricted to count data following Poisson distribution. But, as many have observed, count data often exhibit greater variability than that is provided by the Poisson distribution. In the presence of extra-Poisson variation, one could use regression models based on negative binomial distribution (Tu & Piegorsch, 2003). This research can be extended by linking the negative binomial mean $\mu(x)$ to smooth function $s(x)$. The testing procedure associated with the extended model is currently under investigation.

References

Cameron, A. C. & Trivedi, P. K. (1998). *Regression analysis of count data*. New York: Cambridge University Press.

de Boor, C. (1978). *A practical guide to splines*. New York: Springer-Verlag.

Eilers, P. H. C. & Marx, B. D. (1996). Flexible smoothing with B-splines and penalties. *Statistical Science, 11*, 89-121.

Hastie, T. J. & Tibshirani, R. J. (1990). *Generalized additive models*. New York: Chapman & Hall.

Hope, C., Wu, J., Tu, W., Young, J., & Murray, M. D. (2004). Association of medication knowledge, skill and emergency department visits in older adults with heart failures. *American Journal of Health-System Pharmacy, 61*, 2043-2049.

Lindstrom, M. (1999). Penalized estimation of free-knot splines. *Journal of Computational and Graphical Statistics, 8*, 333-352.

McCullagh, P. & Nelder, J. A. (1989). *Generalized linear model* (2nd ed.). New York: Chapman & Hall.

Ruppert, D. (2002). Selecting the number of knots for penalized splines. *Journal of Computational and Graphical Statistics, 11*, 735-757.

Ruppert, D., Wand, M. P. & Carroll, R. J. (2003). *Semiparametric regression*. New York: Cambridge University Press.

Tu, W., Morris, A.B., Li, J., Wu, J., Young, J., Brater, D. C. & Murray, M. D. (2005). Association between adherence measurements of metoprolol and healthcare utilization in older patients with heart failure. *Clinical Pharmacology and Therapeutics, 77*, 189-201.

Tu, W. & Piegorsch, W. W. (2003). Empirical Bayes analysis of a hierarchical Poisson regression model. *Journal of Statistical Planning and Inference, 111*(1-2), 235-248.